Journal of Computer and Communications, 2015, 3, 196-202
Published Online May 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.35025

Hadoop-Based Similarity Computation System for Composed Documents

Xiaoming Zhang, Zhipeng Qin, Xuwei Liu, Qianyun Hou, Baishuang Zhang, Jie Wu

Department of Computer, Beijing Institute of Petrochemical Technology, Beijing, China
Email: zhangxiaoming@bipt.edu.cn

Received April 2015

Abstract

There exist a large number of composed documents in universities in the teaching process. Most of them are required to check the similarity for validation. A kind of similarity computation system is constructed for composed documents with images and text information. Firstly, each document is split and outputs two parts as images and text information. Then, these documents are compared by computing the similarities of images and text contents independently. Through Hadoop system, the text contents are easily and quickly separated. Experimental results show that the proposed system is efficient and practical.

Keywords

Similarity Computation, Composed Documents, Map Reduce, System Integration

1. Introduction

Document similarity computation is a hot research topic in information retrieval and it is a key issue for automatic document categorization, clustering analysis, fuzzy query and question answering. At present, it aims mainly to improve the accuracy and the efficiency with approaches such as the method based on vector space model [1], the method based on Map-Reduce model [2]. The cloud computing platform with parallel processing ability system, such as Hadoop, is recommended to process large-scale document collection. However, there exist a large number of composed documents in universities with lots of images, tables and text information. These documents may be copied or renewed by some students in the teaching process. This will lead to extra work for the teachers to check the duplication problems. However, the above computing methods are only for text information. They are not suitable directly for these composed documents. On the other hand, it will spend much time to rewrite all the computing tasks in Hadoop system. With the development of data integration, many existed software components are designed easily integrated for computation in data level.

In this paper, we design an integrated system for composed document similarity computation with Hadoop platform and outer program interface. The main works we have done include three aspects: 1) the integration system design solution and its flow-chart; 2) the adopted approaches including documents splitting, image similarity computation with image processing method and text similarity computation using Map Reduce computation model in Hadoop platform; 3) carrying out some related experiments to prove the effectiveness of the ap-

How to cite this paper: Zhang, X.M., Qin, Z.P., Liu, X.W., Hou, Q.Y., Zhang, B.S. and Wu, J. (2015) Hadoop-Based Similarity Computation System for Composed Documents. *Journal of Computer and Communications*, **3**, 196-202.
http://dx.doi.org/10.4236/jcc.2015.35025

proach and system we presented.

2. Integration System Design

A kind of integrated system is presented here in **Figure 1**. It is based on data integration technology for several software systems.

2.1. Document Splitting

The system is designed to process complicated documents embedding images and text information. For the reason, the famous document format, *.doc in Microsoft Word, is chose as the analysis target. All the images in one document will be drawn automatically to be stored into a file folder in the operation file system.

2.2. Map-Reduce Technology

Map-Reduce is a framework for processing huge dataset on certain kinds of distributive problems using a large number of computers, and it is firstly presented by Google and used in Google clouding computing platform. There are also lots of algorithms solving huge dataset based on Map Reduce computation model. Map-Reduce has the ability to increase the computation performance of computer clusters which are composed of PC, and it can solve the problem that a single PC cannot process huge dataset for its limited processor and storage resource. In this framework, the procedure of processing huge dataset could be divided into two steps: Map and Reduce. In each step, it has (key, value) pairs as input, and generates (key, value) pair as output. Therefore, the technology is adopted for text document similarity computation.

3. Design of Similarity Computing Approach

After the document splitting process, its images and text information can be compared independently with other documents one by one.

3.1. Image Similarity Comparing Approach

In order to decrease the computation complexity, each image is processed as fingerprint for comparison, as shown in **Figure 2**.

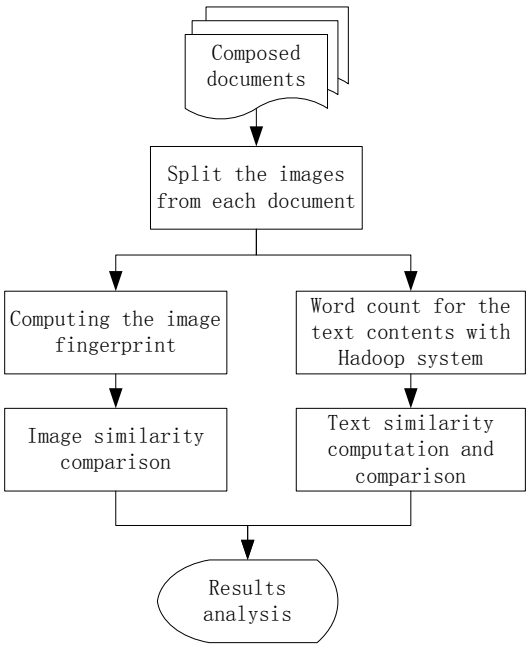

Figure 1. Flow-chart of composed document similarity analysis.

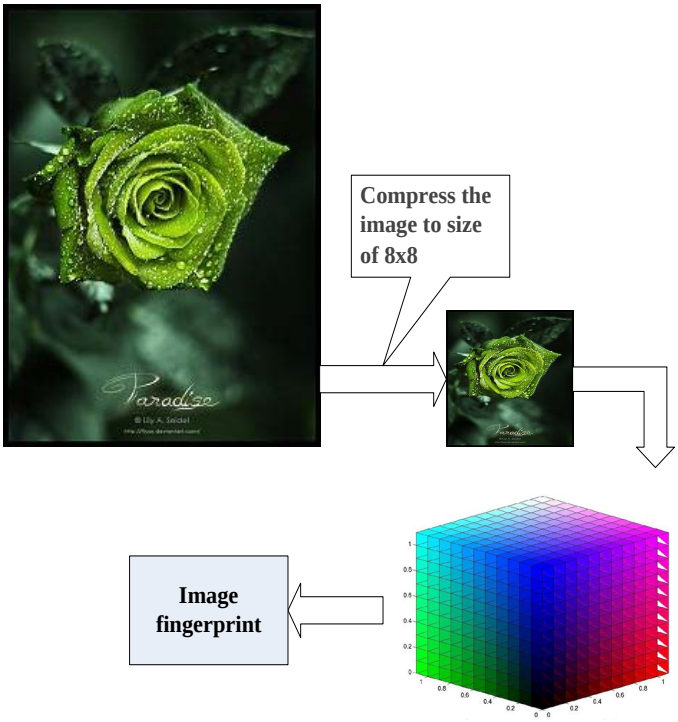

Figure 2. Process of computing image fingerprint.

The procedure of computation is stated as below:
1) Input one of the images in a document.
2) Input another one of images in new document.
3) Compute their fingerprints from the two images as **Figure 2**.
4) Compute their similarity.
5) Go to 2) for inner loop.
6) Go to 1) for outer loop.
7) Output the computation results.

3.2. Document Vector Space Model

Document vector space model is an effective model to perform document similarity computation. Its core idea is firstly to extract feature words $\left(T_i\left(1 \le i \le n\right)\right)$ from document d_k, then d_k could be considered to be composed of feature words group $\left(T_1, T_2, ..., T_n\right)$. For every T_i, it can be assigned certain weight $W_i\left(1 \le i \le n\right)$ according to its significance relative to the document, and then d_k can be represented by the n dimensional vector $V_{dk} = \left(W_1, W_2, ..., W_n\right)$, which contains the value of every feature word's weight. The weight can be computed according to two parameters below: 1) term frequency (TF), which denotes the times of T_i occurs in the document d_j; 2) inverse document frequency (IDF) $\log\left(N / n_i\right)$, where N denotes the number of documents collection and n_i denotes the number of documents which contain T_i. So $W_{i,j}$ which denotes T_i's weight in document d_j can be computed according to the formula as follows:

$$W_{i,j} = f_{i,j} \cdot \log \frac{N}{n_i} \tag{1}$$

$W_{i,j}$ is also be called TF-IDF weight. Based on the document vector space model, the similarity $Sim\left(d_k, d_l\right)$ of document d_k and d_l can be computed according to the cosine value $\cos\left(d_k, d_l\right)$ of vectors space angle. The value of $\cos\left(d_k, d_l\right)$ is proportional to the similarity. If its value is smaller then the similarity between d_k and d_l is lower, otherwise it is contrary. The domain of $\log\left(N / n_i\right)$ is [0, 1], 0 denotes d_k and d_l are ab-

solutely different and 1 denotes they are the same. The computation formula of $\cos(d_k, d_l)$ as follows:

$$\cos(d_k, d_l) = \frac{\sum_{i=1}^{|T|} w_{i,k} \times w_{i,l}}{\sqrt{\sum_{i=1}^{|T|} w_{i,k}^2 \times \sum_{i=1}^{|T|} w_{i,l}^2}} \quad (2)$$

3.3. Text Similarity Computation Approach

As the Map-Reduce model, all the text documents are input to the Hadoop system to operate the Word Count program. The word separating results are stored and return back for text similarity computation. The Map-Reduce computation model can help us to realize the parallel processing.

Next, the output results are transformed to the similarity computing program. Here, all the text information will be calculated and compared using Equation (2). The computation functions are implemented as **Figure 3**.

4. Experimental Analysis

The document format is chosen for *.doc type. Firstly, each document is split to form two parts: one is group with images, and the other is text information. This process is implemented by C# programming. Then, these images are compared for similarity for all documents.

4.1. Document Analysis with Hadoop System

The separated text part from the original document is organized as another text file. Lots of such similar files are input to the Hadoop system for word counting by Map Reduce method. The Hadoop system with version 1.2.1 is adopted under Centos 6.5 operating system. Three computing nodes and one managing node are applied to form the basic Hadoop computing platform. Here, the Word Counting program in Hadoop system is used for all the text files.

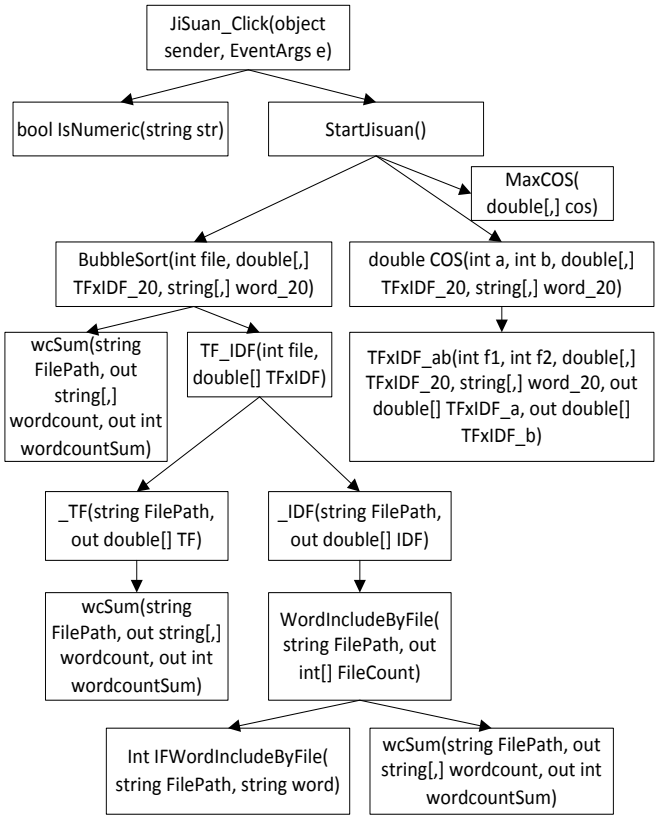

Figure 3. Flow-chart of text similarity computation.

4.2. Image Similarity Comparison

The operation program is implemented using Java, as shown in **Figure 4**. After two images are chosen, their similarity can be quickly obtained and displayed as number.

4.3. Text Similarity Computation

There are ten documents chosen for Map-Reduce computation and comparison. After the WordCount process is finished, the separated words in each document are input to the next program for text similarity computation. The main operation interface is shown in **Figure 5**. The results of similarity are shown in **Figure 6** as matrix.

With the number increasing of words for text similarity, the computation time and biggest similarity show stable, as shown in **Figure 7**. It means that only a small number of words output by Hadoop are necessary for text similarity computation.

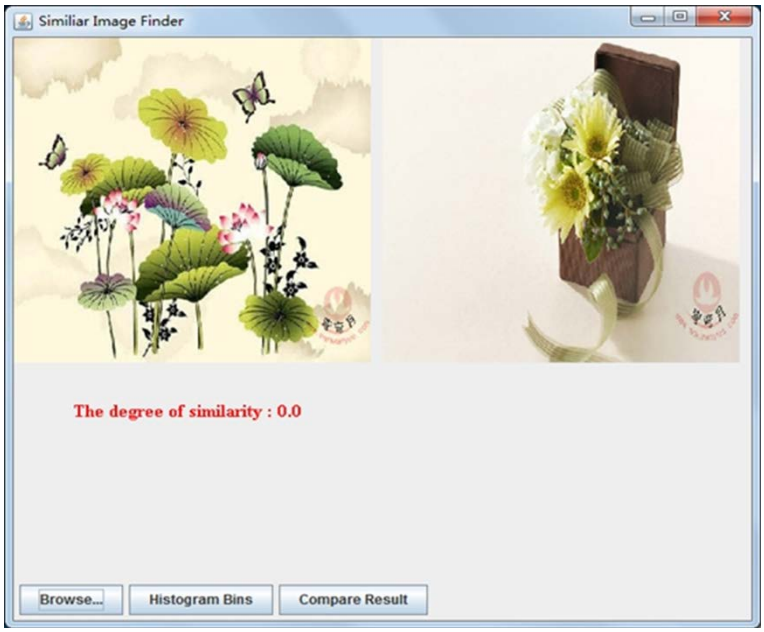

Figure 4. Program operation result for image similarity.

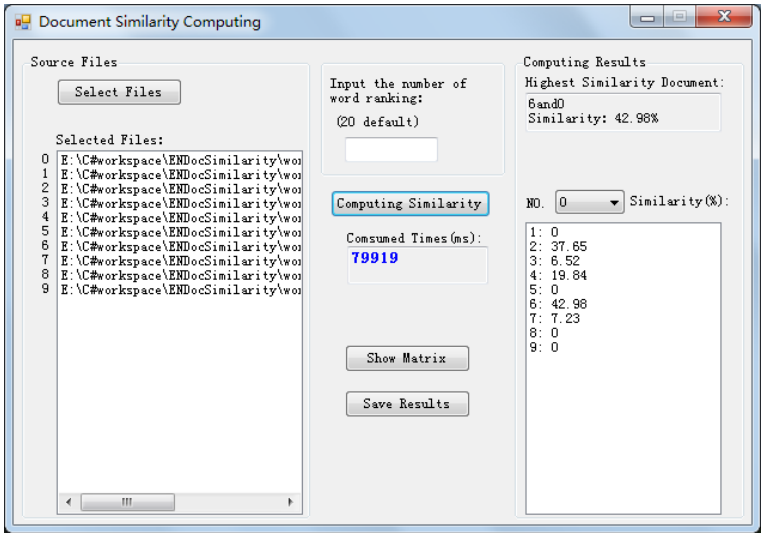

Figure 5. Program operation process for text similarity.

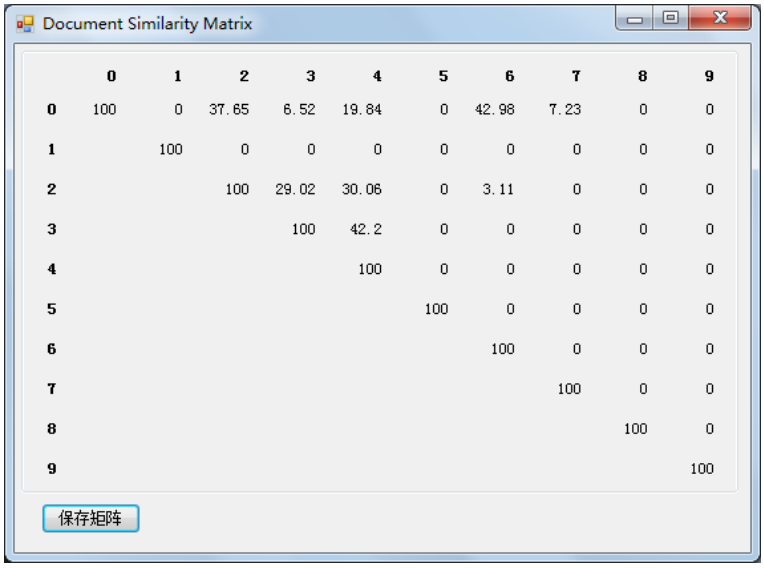

Figure 6. Text similarity presentation as matrix.

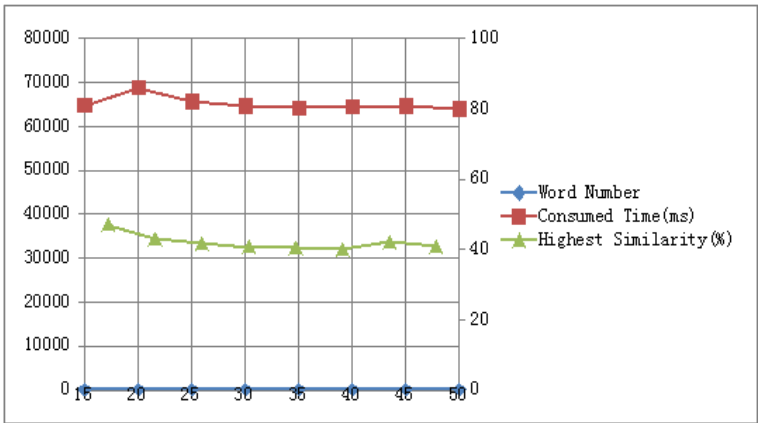

Figure 7. Computation time under highest similarity.

5. Conclusion

Through system integration technology, the composed documents with images and text information can be easily implemented for their similarity. Especially for the text word counting process, the Hadoop system is adopted with its Map-Reduce model. The approaches of image similarity and text similarity show that the proposed integration system is efficient and practical.

Acknowledgements

The work is financially supported by Beijing Institute of Petrochemical Technology with Projects of BIPT-POPME-2015 and the Beijing University Student Scientific Research Plan Project of No. 2014J00099.

References

[1] Mao, E., Wesley, P. and Chu, W. (2007) The Phrase Based Vector Space Model for Automatic Retrieval of Free-Document Medical Documents. Data & Knowledge Engineering, **1**.

[2] He, C.B., Tang, Y. and Tang, F.Y. (2011) Large-Scale Document Similarity Computation Based on Cloud Computing Platform. 2011 6*th International Conference on Pervasive Computing and Applications* (*ICPCA*).

[3] Li, L.N., Li, C.P. and Chen, H. (2013) Map Reduce-Based SimRank Computation and Its Application. 2013 *IEEE In-*

ternational Congress on Big Data.

[4] Baraglia, R., Morales, G.F. and Lucchese, C. (2010) Document Similarity Self-Join with MapReduce. 2010 *IEEE International Conference on Data Mining*. http://dx.doi.org/10.1109/ICDM.2010.70

[5] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 1. http://dx.doi.org/10.1145/1327452.1327492